MELODY EXTRACTION BY MEANS OF A SOURCE-FILTER MODEL AND PITCH CONTOUR CHARACTERIZATION (MIREX 2015)

Juan J. Bosch and Emilia Gómez
Music Technology Group, Universitat Pompeu Fabra, Barcelona, Spain
juan.bosch@upf.edu, emilia.gomez@upf.edu

ABSTRACT

This abstract presents our submission to the MIREX 2015 melody extraction task, whose goal is the identification of the melody pitch sequence from polyphonic musical audio. Our approach combines a source-filter model with the characterisation and analysis of pitch contours. The proposed method obtained the highest overall accuracy on several datasets among the algorithms participating in this year's competition.

1. INTRODUCTION

In our research conducted within the PHENICX project [5], we have evaluated how different melody extraction algorithms perform in the context of symphonic music. The initial step of the evaluation process was to create an annotated collection of symphonic music (ORCHSET). With that purpose, we collected recordings of people singing along with the music. After an analysis of agreement [2], we selected the excerpts in which the annotators sung the same notes and annotated them. The melody in this repertoire is not played by a single instrument, but usually instrument sections which often alternate, and sometimes is not energetically predominant. This poses many challenges to state-of-the-art algorithms, whose accuracy is generally much lower when dealing with such data [2]. The highest melody extraction accuracies are obtained using a source-filter model [4].

In previous MIREX evaluation campaigns, one of the best performing algorithms in terms of overall accuracy is [6]. This approach computes a salience function based on harmonic summation, and then creates and characterises pitch contours for melody tracking. Voicing detection is one of the strong aspects of this method, even though there is still room for improvement since timbre information is not exploited. While this approach works specially well in vocal music, results obtained in MedleyDB dataset [1] showed a drop of 19 percentage points when comparing the overall accuracy obtained in vocal vs. instrumental pieces. In more complex scenarios such as symphonic music, pitch estimation accuracy is degraded, partially due to the simple salience function. This method seems not to be able to cope with the high spectral density, and strong accompaniment.

The motivation behind the present MIREX submission is to combine the estimation accuracy obtained with a generative approach based on a source/filter model with the benefits of using pitch contour characteristics for pitch tracking. Such approach could potentially have good results for relatively simple vocal and instrumental music, as well as in more complex scenarios.

2. METHOD

The proposed method combines a salience function based on a source-filter model (SIMM) [3,4] with a method based on pitch contour characterisation [6].

The spectrogram of a musical audio signal is modelled as the sum of the leading voice and accompaniment. The leading voice is modelled with a source-filter model, and the accompaniment is modelled with a Non-negative Matrix Factorization (NMF). Candidate melody pitch contours are created from the computed salience function, by grouping pitch sequences using auditory streaming cues. Finally, the melody is estimated using contour characteristics and smoothness constraints. We consider a frequency range between 55 Hz and 1760 Hz, and estimate melody pitch values every T=0.01s.

2.1 Pitch Salience Estimation

We model the spectrum X of the signal as the lead instrument plus accompaniment $\hat{X} = \hat{X}_v + \hat{X}_m$. The lead instrument is modelled as: $\hat{X}_v = X_\Phi \circ X_{f_0}$, where X_{f_0} corresponds to the source, X_Φ to the filter, and the symbol $\circ$ denotes the Hadamard product. Both source and filter are decomposed into basis and gains matrices as $X_{f_0} = W_{f_0} H_{f_0}$ and $X_\Phi = W_\Gamma H_\Gamma H_\Phi$ respectively. H_{f_0} corresponds to the pitch activations of the source, and can also be understood as a representation of pitch salience [3]. The accompaniment is modelled as: $\hat{X}_m = \hat{W}_m \hat{H}_m$, leading to Eqn. (1).

$$X \approx \hat{X} = (W_\Gamma H_\Gamma H_\Phi) \circ (W_{f_0} H_{f_0}) + W_m H_m \quad (1)$$

Several parameters need to be specified: the number of bins per semitone (U_{st}), the number of possible elements of the accompaniment (R), the number of atomic filters in W_Γ (K), and the maximum number of iterations (N_{iter}).

Parameter estimation is based on Maximum-Likelihood, with a multiplicative gradient method [4]. In each iteration the parameters are updated in the following order: H_{f_0}, H_Φ, H_M, W_Φ and W_M. The computed salience function is then adapted in order to have an effective pitch contour creation, as will be further detailed in an upcoming publication.

2.2 Pitch Contour Estimation and Melody Selection

From the lead enhanced salience function, we create melody pitch contour candidates by grouping sequences of salience peaks which are continuous in time and pitch, as performed in [6] [1] . Created contours are characterised by a set of features: pitch (mean and deviation), salience (mean, standard deviation), total salience and length. Finally, three further steps are conducted: voicing detection, octave error minimisation (pitch outlier removal), and final melody selection. Previously calculated characteristics are used in this stage to filter out non-melody contours.

3. RESULTS

The evaluation methodology in MIREX compares the sequence of pitches estimated as melody against the ground truth pitch sequence, and focuses on both voicing detection and pitch estimation itself. An algorithm may report an estimated melody pitch even for a frame which is considered unvoiced. This allows the evaluation of voicing and pitch estimation separately. Voicing detection is evaluated using voicing recall and voicing false alarm rates. Pitch estimation is evaluated with Raw Pitch Accuracy (RPA), which is the proportion of melody frames in the ground truth for which the estimation is considered correct (within half a semitone of the ground truth). Raw Chroma Accuracy (RCA) is also a measure of pitch accuracy, in which both estimated and ground truth pitches are mapped into one octave, thus ignoring the commonly found octave errors. Finally, Overall Accuracy (OA) measures the proportion of frames that were correctly labelled in terms of both pitch and voicing. Further details about the metrics can be found in [7]. The evaluation in MIREX 2015 has been performed in five datasets, which contain: north indian vocal classical music (INDIAN), pop, rock, jazz, Rock, R&B, solo classical piano (ADC04, MIREX05), karaoke singing of chinese songs with synthetic accompaniment (MIREX09), and finally our dataset containing symphonic music recordings (ORCHSET) which was used in MIREX this year for the first time.

Table 1 shows the result obtained with three pitch estimation metrics on the symphonic music dataset (ORCHSET), and Table 2 show overall accuracy results on four different datasets in MIREX. The best results in comparison to the rest of the approaches were obtained on the orchestral music dataset. The raw pitch accuracy obtained by our method (BG1) reaches 0.66, which is in relative terms about 86.7% higher than the second best approach (RPA = 0.35). These results show that symphonic music is very challenging for state of the art melody extraction methods, which are generally tailored for vocal music.

[1] http://essentia.upf.edu/

Algorithm	OA	RPA	RCA
BG1	**0.5709**	**0.6615**	**0.8050**
ZCY2	0.3401	0.3543	0.5915
ZCY1	0.3400	0.3542	0.5915
IIY2	0.3273	0.3557	0.6794
IIY1	0.2571	0.2807	0.6552
FYJ1	0.1390	0.1840	0.4517
FYJ4	0.1344	0.1960	0.4582
FYJ2	0.1208	0.2071	0.4714
FYJ3	0.1190	0.1989	0.4362

Table 1. Evaluation results for the Orchset dataset, ordered by Overall Accuracy. Results provided for: RPA: Raw Pitch accuracy, RCA: Raw Chroma accuracy and OAC: Overall Accuracy. Bold font indicates the maximum value for each metric. The proposed method is BG1.

Algorithm	INDIAN	ADC04	MIREX05	MIREX09
BG1	**0.7407**	**0.6930**	0.6274	0.5397
FYJ1	0.6768	0.6007	0.5852	0.7613
FYJ2	0.6897	0.5617	0.5436	0.7467
FYJ3	0.6906	0.5561	0.5441	0.7442
FYJ4	0.6897	0.6169	0.5619	**0.7622**
IIY1	0.6962	0.5843	0.6074	0.6627
IIY2	0.7034	0.6348	**0.6549**	0.6807
ZCY1	0.5048	0.6062	0.4563	0.4623
ZCY2	0.5048	0.6024	0.4563	0.4618

Table 2. Overall accuracy results for 4 different datasets. Bold font indicates the highest values for each dataset.

The proposed method also obtains the best overall accuracy results in INDIAN and ADC04, and second best results on MIREX05, which shows that not only is our method adequate for instrumental music, but also for vocal music from several music genres. Results are worse in karaoke recordings of chinese songs, partially because we did not train our algorithm for this kind of data. In fact, the value of the parameters used for pitch contour creation and melody selection has a very important role in the final accuracy of the method. It is important to note that we only submitted one version of the algorithm, with the parameters tuned to provide a good accuracy on ORCHSET, and the development sets of ADC04 and MIREX05 datasets. Given the heterogeneity of the datasets, it is possible to improve the obtained results with a finer tuning of the parameters, adapted to the characteristics of each of them.

4. ACKNOWLEDGEMENTS

We would like to thank the IMIRSEL team for running MIREX, and the Digital Media Research Center at the Korea Electronics Technology Institute (KETI) team for leading the audio melody extraction task. The research leading to these results is supported by the European Union Seventh Framework Programme FP7 (2007-2013) through the PHENICX project (grant agreement no. 601166).

5. REFERENCES

[1] R. Bittner, J. Salamon, M. Tierney, M. Mauch, C. Cannam, and J. Bello. Medleydb: a multitrack dataset for annotation-intensive mir research. In *15th International Society for Music Information Retrieval Conference (ISMIR 2014)*, pages 155–160, 2014.

[2] J. Bosch and E. Gómez. Melody extraction in symphonic classical music: a comparative study of mutual agreement between humans and algorithms. In *9th Conference on Interdisciplinary Musicology – CIM14*, Berlin, 04/12/2014 2014.

[3] J. Durrieu, B. David, and G. Richard. A musically motivated mid-level representation for pitch estimation and musical audio source separation. *Sel. Top. Signal Process. IEEE J.*, 5(6):1180–1191, 2011.

[4] J. Durrieu, G. Richard, B. David, and C. Févotte. Source/filter model for unsupervised main melody extraction from polyphonic audio signals. *Audio, Speech, Lang. Process. IEEE Trans.*, 18(3):564–575, 2010.

[5] E. Gómez, M. Grachten, A. Hanjalic, J. Janer, S. Jorda, C. Julia, C. Liem, A. Martorell, M. Schedl, and G. Widmer. Phenicx: Performances as highly enriched and interactive concert experiences. *Open access*, 2013.

[6] J. Salamon and E. Gómez. Melody extraction from polyphonic music signals using pitch contour characteristics. *IEEE Trans. Audio. Speech. Lang. Processing*, 20(6):1759–1770, 2012.

[7] J. Salamon, E. Gómez, D. Ellis, and G. Richard. Melody Extraction from Polyphonic Music Signals: Approaches, applications, and challenges. *IEEE Signal Process. Mag.*, 31:118–134, 2014.